DEFINING CORTICAL SULCUS PATTERNS USING PARTIAL CLUSTERING BASED ON BOOTSTRAP AND BAGGING

Zhong Yi Sun[1,2]*, Denis Rivière*[1,2]*, Edouard Duchesnay*[1,2]*,*
Bertrand Thirion[1,2,3]*, Fabrice Poupon*[1,2] *and Jean-François Mangin*[1,2]

[1] Neurospin, I2BM, CEA, Gif-sur-Yvette, France
[2] Institut Fédératif de Recherche 49, France
[3] INRIA Futur, Gif-sur-Yvette, France

ABSTRACT

The cortical folding patterns are very different from one individual to another. Here we try to find folding patterns automatically using large-scale datasets by non-supervised clustering analysis. The sulci of each brain are detected and identified using the brainVISA open software. The 3D moment invariants are calculated and used as the shape descriptors of the sulci identified. A partial clustering algorithm using bootstrap sampling and bagging (PCBB) is devised for cortical pattern mining. Partial clusters are found using a modified hierarchical clustering method constrained by an objective function which looks for the most compact and dissimilar clusters. Bagging is used to increase stability. Experiments on simulated and real datasets are used to demonstrate the strength and stability of this algorithm compared to other standard approaches. Some cortical patterns are found using our method. In particular, the patterns found for the left cingulate sulcus are consistent with the patterns described in the atlas of Ono.

Index Terms— clustering, patterns, sulcus, morphometry

1. INTRODUCTION

The most striking feature when we look at the brain is how convoluted its surface is. Not like other organ such as the heart and the kidney, the cerebral cortex is full of folds, the larger and more stable ones are called the sulci. The cortex folding pattern of each individual is somehow consistent yet very different [1]. A naming system has been developed to name the sulci and the gyri of the brain, but to take one individual brain and try to label the sulci is a very difficult task, even for experienced anatomist, due to the variability of the folding patterns.

Whether these individual differences in folding patterns are related to differences in various skills and functional capabilities is largely unknown. The most detailed description of the sulcus variability has been proposed in the atlas of Ono [2]. This atlas is based on twenty different brains. For each sulcus, the authors propose a list of possible patterns and their frequencies. Our goal is to do the same type of cortical pattern analysis, automatically and on a large scale, with the aid of computers.

Clustering is traditionally used for exploratory data analysis. The data points are assigned to different clusters, such that the members of the same cluster are as similar as possible, while members of different clusters are as different as possible [3]. We designed a Partial hierarchical Clustering method using Bootstrap and Bagging (PCBB), to mine for the patterns. Simulated datasets are used to compare the PCBB algorithm with k-medoid and model-based algorithms, real datasets are used to validate the quality and stability of this algorithm. Some cortical patterns found by the PCBB algorithm are presented.

2. PCBB ALGORITHM

The 3D moment invariants are used as the shape descriptors of the cortical folds. Some investigations were made to ensure that the set of moment invariants used are good shape descriptors [4, 5]. Two aspects were verified. First, the moment invariants vary smoothly in the shape space of the cortical sulci. Second, the moment invariants of different sulci are well separated from each other. The sulci of the brains are detected and labelled automatically by brainVISA software [6]. The 3D moment invariants of each sulcus is then calculated and used as input to the PCBB clustering algorithm. The steps of the PCBB algorithm are described below.

Step 1: Agglomerative hierarchical clustering is performed [7, 5]; the agglomeration process is guided by an objective function:

$$R = \frac{\Sigma \text{ compactness of the clusters formed}}{\Sigma \text{ distance among the clusters formed}} \qquad (1)$$

In each step of the agglomeration process, R is calculated. The p-value of the clusters formed at each step is then estimated by a parametric sampling process [8]. Simulated distributions are generated using the covariance matrix of the real data. The same clustering process is applied to these simu-

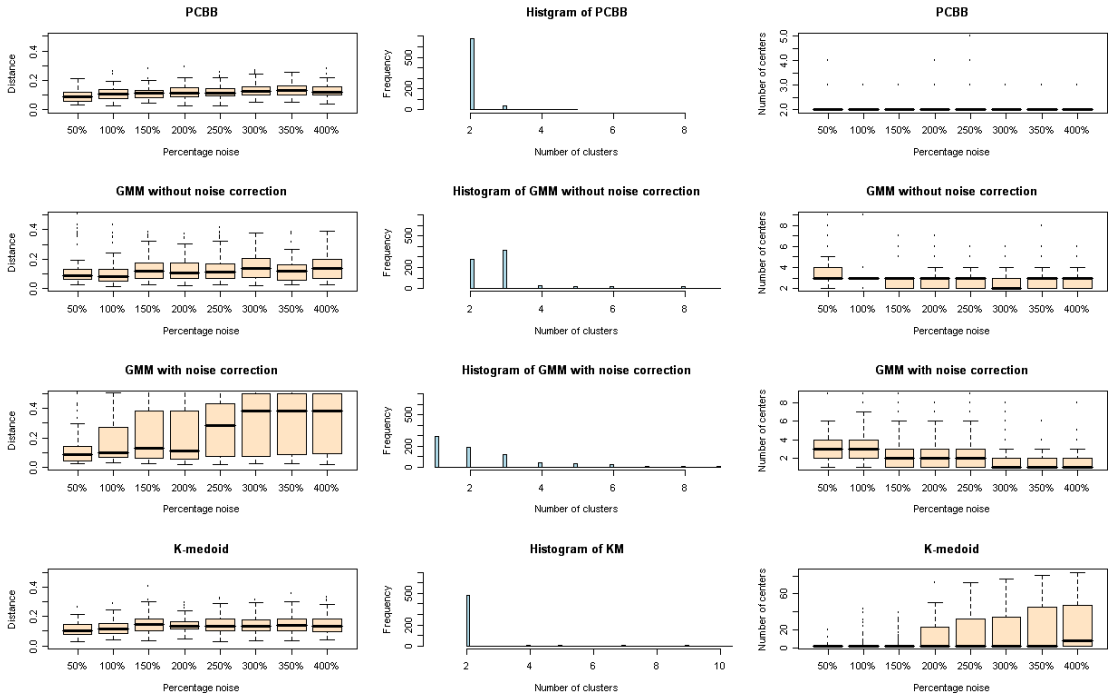

Fig. 1. *The first column shows the boxplots of the distance of the two closest cluster centres found by the algorithm to the real centres. (In the box plot, within the box is the data from the first to the third quartile, the dark line inside the box represents the median. Below the box shows the line of the minimum, above the box the line of the maximum, the outliers are shown as dots.) The x-axis shows the eight simulated datasets accumulated across the pairs of sulci, with percentages of noise from 50 to 400 percents. The results of the PCBB method is shown on the first row, the results of GMM without noise correction are shown on the second row, the results of GMM with noise correction are shown on the third row, and the results of k-medoid are shown on the last row. The second column shows the histogram of the distributions of the number of centres by the three algorithms, the third column shows the boxplot of the number of centres for the eight different datasets.*

lated datasets and p-value is estimated by counting the number of times the simulated data have a better R score than the real data. Finally the clusters with the best p-value are chosen as the salient points for the next step of the algorithm.

The goal of this step is to estimate the number of clusters and their size automatically. Notice that the clustering is "partial", not all data points are assigned to clusters. The goal here is to extract the most interesting sample points that might contain strong and significant patterns. We are not trying to assign each point to a pattern. Note also that the p-values estimated here are only used in the ranking system to pick out the most interesting clusters; they are not used to perform statistical tests.

Step 2: The process described in step 1 is performed many times on the bootstrap datasets of the original data. A repertoire of salient points is identified to form a new dataset. A simple K-medoid algorithm (the number of clusters being estimated by the standard PAM criterion) is then used to find the final clusters [7]. This step is using the idea of bagging [3]; the goal is to overcome the instability of the clusters found in the first step. The assumption is that the first step of the clustering on the bootstrap samples gives the strongest clusters and eliminates most of the noise. So in this second step a relatively simple clustering algorithm is sufficient to identify the final clusters.

3. RESULTS

3.1. The experiments on simulated datasets: comparison of algorithms

To evaluate the performance of the PCBB algorithm, we perform some experiments on simulated datasets. The procedure and results are presented below. The dataset we use as a model for generating the simulated dataset is a real dataset made up of 36 brains, where each sulcus has been reliably labelled manually by a neuroanatomist. This dataset is used to train the sulcus recognition system of brainVISA. We chose the moment invariant data of the ten biggest sulci of each hemisphere for further analysis. To generate the simulated dataset to evaluate performance, we generated datasets using the mean and covariance matrix of any pair of the ten sulci.

The simulated datasets are generated as follows: take the mean and covariance matrix of any pair of sulci, generate a new dataset using these same parameters. This gives a dataset

with two known clusters. Then a series of noisy datasets are generated by adding $50, 100, 150, 200, 250, 300, 350$ and 400% of noise to the new dataset. The noise added follows a Poisson distribution, within the min and max value of the original dataset in their respective dimensions. The mean and covariance matrix of the real sulci are used to keep the simulated data close to the distribution of the real data.

K-medoid-style clustering algorithm, the Gaussian mixture modelling algorithm and the PCBB algorithm are run on these simulated datasets. The results are evaluated in terms of the number of clusters found and how far are the cluster centres found from the real centres.

K-medoid is a variant of the k-means algorithm that uses the optimal representative sample (medoid) of its cluster. It is considered more robust with respect to outliers than k-means [7]. The standard PAM criterion is used to find the optimal number of clusters [7].

The Gaussian mixture model-based (GMM) clustering involves first fitting a mixture model, usually by the expectation-maximization (EM) algorithm, and then the posterior probability of each mixture component is computed given a data point [3]. Some success has been shown using the Bayesian Information Criterion (BIC) to choose the right number of components. However, in general, equating a component of GMM with a cluster is questionable [9]. In our experiments, we use the Mclust toolbox from R to run the GMM algorithm [10, 11]. Mclust is a state-of-the-art mixture-model-based clustering tool. We did two GMM runs for each dataset. The first run allows the algorithm to optimally select the structure of the covariance matrices using the standard BIC criterion [3], but without the initialization of the proportion of noise as a prior. In the second run the real proportion of noise in the dataset is given as a prior to the algorithm.

Two comparisons are made to access the quality of the clustering. First, the distance of the clusters centres found to the real centres are measured. When there are more cluster centres found by the algorithm than the real centres, only the two clusters closest to the real centres are taken into consideration. Second, the numbers of cluster centres found by the algorithm are compared for each simulated dataset. The result is shown in Fig. 1.

Results show that the PCBB algorithm is comparable to k-medoid and to the GMM algorithm in terms of locating the centres of clusters. However, in terms of estimating the number of clusters, PCBB is a lot more accurate and stable than the other algorithms, with increasing number of noise in the data. Feeding the GMM a percentage of noise during initialization does not seem to help the performance. The result shows that PCBB is more robust than GMM and k-medoid for the particular problem of finding cortical patterns.

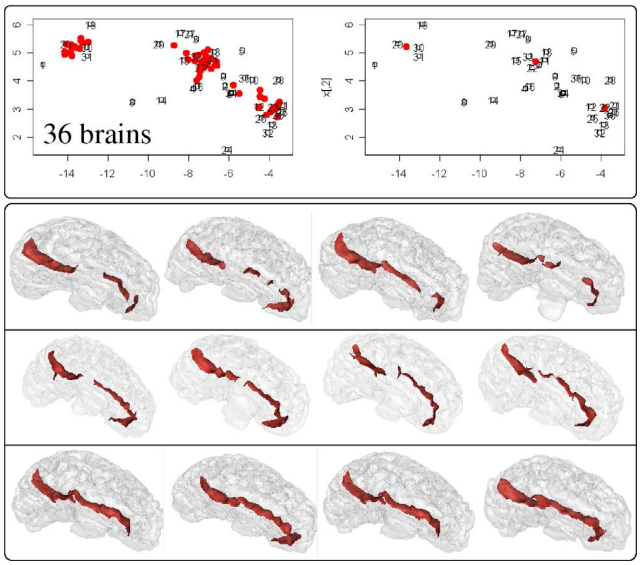

Fig. 2. *The top-left image shows the salient points found with 100 bootstrap samples. The top-right image shows the final cluster centres found PCBB. The X and Y-axis are the first and second dimensions of the PCA. The bottom image shows the shapes of the cingulate sulcus of the three corresponding clusters (from left to right in the PCA).*

3.2. The experiments on real datasets

The clustering experiment is run on the 36 brain dataset used for generating the simulated data. The clusters found for one of the sulci with the strongest patterns can be observed in Figure 2. Here we observe the forms of the patterns found. The first cluster has a pattern with an anterior interruption, the second cluster has a pattern with a posterior interruption, and the third cluster appears to be continuous. These patterns found are consistent with those described in the atlas of Ono, which stated that around 60% of the instances of the cingulated sulcus have no interruption, around 24% have two segments with a posterior interruption or an anterior interruption, and around 16% are divided into three segments. It should be noted that the size of the database used prevents the detection of rare patterns. Therefore, larger and more comprehensive databases will be required to achieve a more exhaustive pattern search.

Next the PCBB algorithm is performed on two other real datasets. Moment invariant data of the left cingulate sulcus is used like above. The clusters found are shown in Figure 3. Here two large real datasets are used containing respectively 148 and 176 brains. The sulci are automatically identified and labelled by brainVISA. We observe that the PCBB algorithm is stable over different datasets. The bagging procedure used in the second part of the algorithm helps to achieve greater stability.

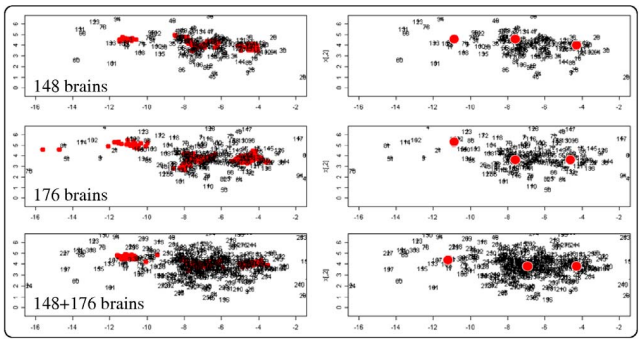

Fig. 3. *The first row shows one dataset, the second row shows another dataset, the third row shows the dataset composed by mixing the data of the two datasets. The charts use the first two axes of the PCA performed on the merged set. The first column shows the salient points found by the first part of PCBB, using 100 bootstrap samples. The second column shows the final cluster centres found by the second part of PCBB.*

4. DISCUSSION

The strong point of the PCBB algorithm is that it is very robust. Also it makes no assumption that the clusters span the whole data space, as all the division-based clustering algorithms do.

When we think about our problem of finding cortical folding patterns here, we are aware that the cortical folding process is very complicated. It can be considered as a chaotic phenomenon. The final global folding pattern of the brain is the end product of numerous chemical and mechanical forces very well orchestrated throughout the time of brain development [12, 13, 14]. Here we are not trying to model the folding process and explain all the variability we observe in folding patterns. Instead we are trying to identify some typical patterns that might exist in only a part of the population, but are significant and can give us some insight into the folding process and certain pathologies. So partial clustering is more relevant for this particular problem.

If such patterns can be defined, we hypothesize that their relative frequencies could be different in certain patient populations, compare to the normal populations. Developmental pathologies, indeed, could modify the dynamics of the folding process and favour some folding patterns over the others. Consequently, the folding patterns could provide some signatures useful for diagnosis.

As an ongoing work we are investigating other shape descriptors for cortical folds. Possible improvements to the clustering algorithm are investigated as well. Finally, we are looking for ways to look for patterns without the knowledge of the traditional nomenclature of the folds.

5. REFERENCES

[1] W. Welker, *Cerebral Cortex*, vol. 8B, chapter Why does cerebral cortex fissure and fold?, pp. 3–136, Plenum Press, New York, 1988.

[2] M. Ono, S. Kubik, and C. D. Abernathey, *Atlas of the cerebral sulci*, Thieme, New York, 1990.

[3] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*, Wiley-Interscience, 2000.

[4] J.-F. Mangin, F. Poupon, E. Duchesnay, D. Rivière, A. Cachia, D. L. Collins, A. C. Evans, and J. Régis, "Brain morphometry using 3D moment invariants," *Medical Image Analysis*, vol. 8, pp. 187–196, 2004.

[5] Z. Y. Sun, D. Riviè, F. Poupon, J. Régis, and J.-F. Mangin, "Automatic inference of sulcus patterns using 3d moment invariants," in *MICCAI, Brisbane, LNCS-4792, Springer-Verlag*, 2007, pp. 515–522.

[6] D. Rivière, J.-F. Mangin, D. Papadopoulos-Orfanos, J.-M. Martinez, V. Frouin, and J. Régis, "Automatic recognition of cortical sulci of the human brain using a congregation of neural networks," *Medical Image Analysis*, vol. 6, no. 2, pp. 77–92, 2002.

[7] L. Kaufman and P. J. Rousseuw, *Finding groups in data*, Wiley series in probability and statistics, 1990.

[8] P. Good, *Permutation, Parametric, and Bootstrap Tests of Hypotheses*, Springer Series in Statistics, 2004.

[9] S. Ray and B.G.Lindsay, "The topography of multivariate normal mixtures," *Annals of Statistics*, vol. 33, no. 5, pp. 2042–2065, 2005.

[10] C. Fraley and A.E.Raftery, "Model-based clustering, discriminant analysis, and density estimation," *Journal of the American Statistical Association*, vol. 97, pp. 611–631, 2002.

[11] C. Fraley and A.E.Raftery, "Mclust version 3 for r: Normal mixture modeling and model-based clustering," *Technical Report, no.504, Department of Statistics, University of Washington*, 2006.

[12] D. C. Van Essen, "A tension-based theory of morphogenesis and compact wiring in the central nervous system," *Nature*, vol. 385, pp. 313–318, 1997.

[13] T. Fukuchi-Shimogori and E. Grove, "Neocortex patterning by the secreted signaling molecule fgf8," *Science*, vol. 294, 2001.

[14] E. Monuki and K. Walsh, "Mechanisms of cerebral cortical patterning in mice and humans," *Nature neuroscience*, vol. 4, 2001.